DOCUMENT RESUME

ED 475 274 UD 035 615

AUTHOR	Rockoff, Jonah E.
TITLE	The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.
SPONS AGENCY	Harvard Univ., Cambridge, MA. Kennedy School of Government.
PUB DATE	2003-03-00
NOTE	41p.
AVAILABLE FROM	For full text: http://econwpa.wustl.edu:8089/eps/pe/papers/0304/0304002.pdf.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; Elementary Education; English (Second Language); Ethnicity; Mathematics Skills; Reading Skills; Teacher Competencies; *Teacher Influence; *Teaching Experience
IDENTIFIERS	New Jersey

ABSTRACT

In order to provide accurate estimates of how much teachers affect the achievement of their students, this study used panel data covering over a decade of elementary student test scores and teacher assignment in two contiguous New Jersey school districts. The test score data, which spanned the years 1989-1990 to 2000-2001, came from nationally standardized basic skills reading and math tests. Data were also collected on students' gender, ethnicity, special education classification, and English as a Second Language enrollment, as well as on school, grade, and teacher identifiers. The study estimated teacher fixed effects while controlling for fixed student characteristics and classroom specific variables. Data analysis indicated that there were large and statistically significant differences among teachers. A one standard deviation increase in teacher quality raised students' reading and math test scores by approximately .20 and .24 standard deviations, respectively, on a nationally standardized scale. In addition, teaching experience had statistically significant positive effects on reading test scores, controlling for fixed teacher quality. (Contains 25 references.) (SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 475 274

The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data

Jonah E. Rockoff*
March, 2003

Abstract

Teacher quality is widely believed to be important for education, despite little evidence that teachers' credentials matter for student achievement. To accurately measure variation in achievement due to teachers' characteristics—both observable and unobservable—it is essential to identify teacher fixed effects. Unlike previous studies, I use panel data to estimate teacher fixed effects while controlling for fixed student characteristics and classroom specific variables. I find large and statistically significant differences among teachers: a one standard deviation increase in teacher quality raises reading and math test scores by approximately .20 and .24 standard deviations, respectively, on a nationally standardized scale. In addition, teaching experience has statistically significant positive effects on reading test scores, controlling for fixed teacher quality.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. Rockoff
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

*E-mail: rockoff@fas.harvard.edu. I wish to thank Gary Chamberlain, David Ellwood, Caroline Hoxby, Christopher Jencks, and Larry Katz for extensive comments, as well as Raj Chetty, Adam Looney, Sarah Reber, Tara Watson, and seminar participants at Harvard University for many helpful suggestions and discussions. I am most grateful to the local administrators who worked with me to make this project possible. This work was supported by a grant from the Inequality and Social Policy Program at Harvard's Kennedy School of Government.

UD 035 615

1 2 BEST COPY AVAILABLE

1 Introduction

School administrators, parents, and students themselves widely support the notion that teacher quality is vital to student achievement, despite the lack of evidence linking achievement to observable teacher characteristics. Studies that estimate the relation between achievement and teachers' characteristics, including their credentials, have produced little consistent evidence that students perform better when their teachers have more 'desirable' characteristics. This is all the more puzzling because of the potential upward bias in such estimates—teachers with better credentials may be more likely to teach in affluent districts with high performing students.[1]

This has led many observers to conclude that, while teacher quality may be important, variation in teacher quality is driven by characteristics that are difficult or impossible to measure. Therefore, researchers have come to focus on using matched student-teacher data to separate student achievement into a series of "fixed effects," and assigning importance to individuals, teachers, schools, and so on. Researchers who have sought to explain wage determination have followed a similar empirical path; they try to separate industry, occupation, establishment, and individual effects using employee-employer matched data (Abowd and Kramarz, 1999). Despite agreement that the identification of teacher fixed effects is a productive path, this exercise has remained incomplete because of a lack of adequate data. Credible identification of teacher fixed effects requires panel data where students and teachers

[1] Teachers with better credentials, such as experience or selectivity of undergraduate institution, tend to gravitate towards districts with higher salaries (Figlio, 1997).

are observed in multiple years, and this type of data is not readily available to researchers.[2]

A small number of studies have found significant variation in test scores across classrooms within particular schools, even after controlling for student characteristics.[3] In other words, dummy variables identifying students' classrooms seem to be important explanatory variables in regressions of student test scores. Although researchers have associated the significance of classroom dummy variables with variation in teacher quality, other classroom specific factors may also be driving differences among classroom achievement levels. In these studies, teacher effects cannot be separated from other classroom effects because teachers are only observed in one classroom.

In order to provide more accurate estimates of how much teachers affect the achievement of their students, I obtained panel data covering over a decade of student test scores and teacher assignment in two contiguous school districts. The observation of teachers with multiple classrooms allows me to measure teacher fixed effects while including direct controls for a number of classroom specific factors that may systematically influence student test score performance, such as peer achievement and class size. Observation of students' test scores in multiple years allows for the inclusion of student fixed effects, so that variation in student fixed characteristics, such as cognitive ability, does not drive estimated differences in student performance across teachers. In addition, because teachers' experience levels change naturally, the effect of experience on student performance is identified from variation within teachers. This is, to my knowledge, the first study of teacher quality that employs

[2] However, this data is widely collected by both local and state education agencies, and could be used by these institutions for purposes of evaluation. Though this has seldom been done in practice, one prominent example is the Tennessee Value-Added Assessment System, where districts, schools, and teachers are compared based on test score gains averaged over a number of years.

[3] Hanushek (1971), Murnane (1975), and Armor et al. (1976).

these methods.

Estimates of teacher fixed effects from linear regressions of test scores consistently indicate that there are large differences in quality among teachers in this data. A one standard deviation increase in teacher quality raises test scores by approximately .20 standard deviations in reading and .24 standard deviations in math on nationally standardized distributions of achievement. I find that teaching experience significantly raises student test scores in reading subject areas. Reading test scores differ by approximately .20 standard deviations on average between beginning teachers and teachers with ten or more years of experience. Moreover, estimated returns to experience are quite different if teacher fixed effects are omitted from my analysis. This suggests that using variation across teachers to identify experience effects may give biased results due to correlation between teacher fixed effects and teaching experience.

Policymakers have demonstrated their faith in the importance of teachers by greatly increasing funding for programs that aim to improve teacher quality in low performing schools.[4] However, the vast majority of these initiatives focus on rewarding teachers who possess credentials that have not been concretely linked to student performance (e.g. certification, schooling, teacher exam scores). My results support the idea that raising teacher quality is an important way to improve achievement, but suggest that policies may benefit from shifting focus from credentials to performance-based indicators of teacher quality.

This paper is organized as follows: in section two, I provide an overview of previous

[4]The most recent example is the 'No Child Left Behind Act,' which appropriated over $4 billion for training and recruitment of teachers in 2002. This is in addition to various other federal and state initiatives targeting teachers, such as forgiving student loans, easing qualifications for home mortgages, and waiving tuition for teachers' children who enroll in state universities.

literature on the importance of teachers; in section three, I describe the data I collected for this study; in section four, I present my empirical findings; section five concludes.

2 Related Literature

The overwhelming majority of work on teacher quality has examined the relation between teacher characteristics and objective measures of student performance (usually standardized test scores), at the individual, school, or district level. Hanushek (1986) provides an accounting of the results of 147 such studies. With regard to teacher education and teacher experience, he finds, "In a majority of cases, the estimated coefficients are statistically insignificant. Forgetting about statistical significance and just looking at estimated signs does not make much of a case for the importance of these factors either." The lack of any consistent pattern in these results is striking, considering the fact that most schools pay more for teachers with graduate degrees or more experience, suggesting a belief on their part that these factors indicate higher teacher quality. It is even more surprising if one believes that non-random assignment of teachers to schools and/or classrooms could lead to a positive bias in any estimated relation between teachers' credentials and student achievement.

However, these findings should not be taken as strong evidence that teachers do not matter; only that teacher quality may be unrelated to these observable characteristics. A more direct method to address whether teachers matter is to test whether there are statistically significant differences in students' achievement levels caused by persistent differences among their teachers—in other words, to identify teacher fixed effects. Yet only a small number of studies have even approached the problem in this way because the data required to do so is

rarely available to researchers.[5]

Hanushek (1971) was the first to use fixed effects in an analysis of student achievement. He demonstrated that classroom dummy variables have significant explanatory power in regressions of students' test scores—conditional on past achievement—and took this as an indication that teacher effects are important. Classroom dummy variables were similarly found to be significant predictors of test scores in studies by Murnane (1975) and Armor et al. (1976).[6] In addition, both studies found that principals' opinions of teachers had predictive power for student achievement, providing some evidence that teacher quality was driving a significant portion of the variation in achievement across classrooms. Teachers were only observed with one classroom in all three studies, so teacher effects could not be directly separated from classroom effects in their analyses.

Hanushek (1992) again found significant differences among classrooms using data on Black children from the Gary Income Maintenance Experiment.[7] More importantly, some teachers were observed with multiple classrooms, allowing for a direct test of whether teachers, as opposed to other classroom factors, were driving differences in achievement across classrooms. Hanushek's strategy was to test the restriction of equal effects across classrooms with the same teacher. He found the restriction could be rejected, but only due to

[5] Rivkin et al. (2001) take a more subtle approach to estimating teacher quality and deserve mention. Though they cannot match students with their actual teachers, they model a link between teacher turnover and teacher quality that occurs through changes in the variance of test score gains across cohorts. They use this model to construct a lower bound estimate of the contribution of teacher quality to student test scores using data from the Texas Schools Project. They find a one standard deviation increase in teacher quality raises test scores by .11 standard deviations.

[6] All three studies also examined a multitude of teacher characteristics, including education and experience, but none were consistently found to have significant predictive power for student test scores. Hanushek found that performance on a *Quick-Word Test* did correlate well with student achievement, which he interprets as proxying for teacher IQ or ability. Murnane found that children performed better with experienced teachers, with male teachers, and also with teachers of the same race.

[7] He again found little evidence that teacher characteristics matter. It is also worth noting that teacher quality was not the main focus of this paper.

a small number of teachers with widely different measured effects across years. He therefore concluded "the general stability of teacher impacts" provided "additional support for a teacher skill interpretation of differences in classroom performance." However, even if there were significant differences in effects for classes with the same teacher, this may reflect the importance of other classroom factors, and not the insignificance of teachers.

The most credible way to identify teacher effects is to regress test scores on teacher dummy variables when teachers are observed with many classrooms (netting out idiosyncratic variation in classroom performance) and controlling for variation in student characteristics and other classroom specific variables. These options were not possible in previous studies because of data limitations. The data I collected for this study links teachers with their students for a period of up to twelve years, and contains up to five years of annual test scores for all elementary students in a number of schools. I can thus credibly identify teacher fixed effects, and measure the importance of teachers by the magnitude of the difference between high and low quality teachers in my data.

3 The Data

I obtained data on elementary school students and teachers in two contiguous districts from a single county in New Jersey. I will refer to them as districts 'A' and 'B'.[8] Both have multiple elementary schools serving each grade, and, within each school, there are two to seven teachers per grade in any particular year. Elementary school populations in these districts grew considerably over this time period, but the racial composition of the students

[8] For reasons of confidentiality, I refrain from giving any information that could be used to identify these districts.

was stable.[9] The average socioeconomic status of residents in these school districts is above the state median, but considerably below the most affluent districts.[10] In the proportion of students eligible for free/reduced price lunch, these districts fell near the 33rd percentile in the state during the 2000-2001 school year. Spending per pupil that year was slightly above the state average in district A, and slightly below average in district B.

I focus on elementary education for four primary reasons. First, elementary students in these districts are tested in the spring of every year using nationally standardized exams; older students are tested less frequently and use state-based examinations. Second, elementary students remain with a single teacher for most of the school day and receive reading and math instruction from this teacher. I can therefore be confident that a student's current teacher is the person from whom they have received almost all instruction since the last time they were tested. Third, school administrators in these districts claim that, unlike higher grades, elementary school students are not tracked by ability or achievement. In the appendix, I confirm this by showing that students are not systematically grouped by previous achievement levels or by their previous classroom.[11] Fourth, it is very likely that basic skills test scores are related to important economic outcomes, and that basic skills constitute a large part of what elementary school students learn.[12]

[9]Enrollment in both districts grew by over 40% during the period for which I have data. Students in these districts are predominantly White (between 70-80% during this time period), with the remainder made up of relatively equal populations of Black, Hispanic, and Asian students.

[10]School districts in New Jersey are placed into District Factor Groups based on the average socio-economic status of their residents, using a composite index of indicators from the most recent U.S. decennial census.

[11]That is, I demonstrate that dummy variables for students' current classrooms do not have predictive power for students' previous test scores. I also show that actual classroom assignment produces similar mixing of classmates from year to year as one would expect from random assignment.

[12]Concerns over 'teaching to the test' are important to the extent that teachers can raise elementary students' basic skills test scores without actually teaching them basic skills. I regard this as highly unlikely. However, teachers who focus on basic skills may do so at the expense of other valuable skills. I have no way of discerning whether or not this occurs in my data.

The test score data I collected spans the 1989-1990 through 2000-2001 school years.[13] Test scores come from nationally standardized basic skills reading and math tests, and up to four subject area tests were given to students in a given year: Reading Vocabulary, Reading Comprehension, Math Computation and Math Concepts.[14] Data collected from District A comes from students in 1st through 5th grade, and in District B from 2nd through 6th grade. Students' scores are reported on a Normal Curve Equivalent (NCE) scale. NCE scores range from 1 to 99 (with a mean of 50 and standard deviation of 21) and are standardized by grade level.[15] Test makers assert that each NCE point represents an equal increase in test performance, allowing scores to be added, subtracted, or averaged in a more meaningful way than national percentiles. Using national percentiles in my analysis does not noticeably alter the results. Figure 1 shows the distribution of NCE scores in these districts, along with the nationally standardized distribution.[16] Students in these districts score 10-15 NCE points higher on average than the nationwide mean in all subjects. The variance in test score performance within these districts is considerable, though less than the national distribution, and relatively few students score below 30 NCE points.[17]

[13] District B data does not include the 2000-2001 school year.

[14] Both districts administered the Comprehensive Test of Basic Skills (CTBS) at the start of these time period, but switched at some point—District A to the TerraNova CTBS (a revised version of CTBS) and District B to the Metropolitan Achievement Test (MAT). The subtest names are identical across all of these tests, and it is therefore unlikely that the changes reflect a radical shift in the type of material tested or taught to students.

[15] Using scores that are standardized at a particular grade level may be problematic if the distribution of student achievement changes as students grow older. For example, if a change of one NCE point at the 6th grade level represents a much smaller difference in learning than one NCE point at the 1st grade level, then we might want to regard a given amount of variation in 1st grade student performance as representing larger variation in teacher quality than the same variation among 6th graders. I do not attempt to reconcile this possibility in my analysis.

[16] I pool the districts because their distributions of scores are quite similar. Not all 99 scores are possible on every test, so I group NCE scores from 1-9, 10-19, and so on. I simulate the national distribution by taking 20,000 draws from a normal distribution with mean 50 and standard deviation 21, and then grouping them in the same manner.

[17] A small but non-trivial percentage (about 3-6%) of scores in each district are at the maximum possible for the test taken, raising the possibility that censoring of 'true' achievement might affect the results of my

I also gathered information on students' gender, ethnicity, special education classification and ESL enrollment, as well as school, grade, and teacher identifiers. Teacher identifiers are also matched with data on their highest degree earned, teaching experience, and year of birth.[18] As mentioned above, the usefulness of this dataset stems from the observation of both pupils and teachers in multiple years. In both districts, the median student was tested three times—almost one quarter of the students were tested once, and over one quarter were tested five times. The median number of classrooms observed per teacher is six in district A and three in district B. About 18% of teachers in district A and 26% of teachers in district B are only observed with just one classroom of students, but 53% of teachers in district A and 29% of teachers in district B are observed with more than five classrooms of students.

Analyzing the districts separately reveals no marked differences in results or conclusions, and for simplicity I combine them in the results presented below. Because the number of tests administered varies somewhat over grades and years, and because teacher quality may vary by subject, I examine each subject area separately, and then consider to what extent my results differ across them.

---

analysis. I checked for this by performing the main part of my analysis with censored-normal regressions, and the results were not qualitatively different to those presented below. Also, enrolled students who are absent on the day of the test, or change districts earlier in the year, are not observed in the testing data. To see whether the probability of being tested was related to achievement, I use enrollment information available in district B since the 1995-1996 school year. I find no significant relationship between students' previous test scores and their probability of being tested in the following year, both in linear probability and probit regressions.

[18] Information on teachers' education and experience was not available for a small portion of teachers in both districts, and for some teachers I only know their experience teaching in the district. However, for teachers where data is not missing, the vast majority had no previous teaching experience when hired. In any case, omitting teachers with incomplete data from my analysis does not have a noticeable effect on the results.

4 Empirical Results

$$(1)\ A_{isgjt} = \alpha_i + \gamma X_{it} + \theta_j + f(Exp_{jt}) + \eta C_{jt} + \pi_s + \pi_g + \pi_t + \varepsilon_{isgjt}$$

Consider equation 1, which provides a linear specification of the test score of student i in school s and grade g, with teacher j in year t. The test score (A_{isgjt}) is a function of the student's fixed characteristics (α_i), observable time-varying characteristics (X_{it}), a teacher fixed effect (θ_j), teaching experience (Exp_{jt}), observable classroom characteristics (C_{jt}), factors varying across schools, grades, and years (π_s,π_g,π_t), and all other factors that affect test scores (ε_{isgjt}), including measurement error.[19] This model restricts effects to be independent across ages, and assumes no correlation between current inputs and future test scores—zero persistence—except for inputs that span across years, like α_i.[20]

Two issues of collinearity create difficulties in the estimation of equation 1. Experience and year are collinear within teachers (except for a few who leave and return) and grade and year are collinear within students (except for a few who repeat grades).[21] Because of these issues, consistent estimation of teacher fixed effects and experience effects can be achieved only under some identifying assumptions.

The first assumption I make is that additional experience does not affect student test scores after a certain point ($f' = 0$ if $Exp_{jt} > \overline{Exp}$). Under this assumption, year effects

[19] Subscripts for subject area are not included for simplicity. Experience is defined as number of years taught prior to the current year, so that new teachers are considered to have zero experience.

[20] Persistence of effects will bias my estimates of teacher fixed effects if the quality of current inputs and past inputs are correlated, conditional on the other control variables. Because classroom assignment appears similar to random assignment in these districts (see appendix), this source of bias is likely to be unimportant. A simple way to incorporate persistence, used in a number of other studies, is to model test score gains, as opposed to levels. However, this type of model restricts changes in test scores to be perfectly persistent over time, which, if not true, would lead to the same type of bias. Also, test scores gains can be more volatile, since the idiosyncratic factors that affect test score levels will affect gains to twice the extent (Kane and Staiger, 2001).

[21] In these districts, I find 9% of teachers had discontinuous careers, and less than 1% of students repeated grades.

can be separately identified from students whose teachers have experience above the cutoff ($\overline{Exp}$). This restriction is supported by previous research, which suggests that the marginal effect of experience declines quickly, and any gains from experience are made in the first few years of teaching (Rivkin et al., 2001). Moreover, the plausibility of this assumption can be examined by viewing the estimated marginal experience effects at $\overline{Exp}$.[22] My second assumption is that grade effects are zero and can therefore be omitted from the model. This assumption is supported by the fact that test scores are normalized by grade level.[23] Equation 2 incorporates these two identifying assumptions and generalizes the model by including school-year effects (π_{st}).[24]

$$(2)\ A_{isjt} = \alpha_i + \gamma X_{it} + \theta_j + f\left(Exp_{jt}\right) D_{Exp_{jt} \leq \overline{Exp}} + f\left(\overline{Exp}\right) D_{Exp_{jt} > \overline{Exp}} + \eta C_{jt} + \pi_{st} + \varepsilon_{isjt}$$

Table 1 shows results for regressions where $f\left(Exp_{jt}\right)$ is a cubic and $\overline{Exp}$ equals ten years of experience.[25] The time-varying student controls (X_{it}) are dummy variables for being retained or repeating a grade, and the classroom controls (C_{jt}) are class size, the average of classmates' test scores from the previous year, being in a split-level classroom, and being in the lower half of a split level classroom.[26] Because errors are heteroskedastic and possibly

[22] For example, if $f\left(Exp_{jt}\right)$ is estimated as a quadratic, then $f\left(Exp_{jt}\right) = aExp_{jt} + bExp_{jt}$, and one can test whether $a + 2b\overline{Exp} = 0$.

[23] Technically, consistent estimation of teacher fixed effects only requires that grade effects be uncorrelated with teacher assignment, but this is clearly not true—the majority of teachers in these districts do not switch grade levels.

[24] $D_{Exp_{jt} \leq \overline{Exp}}$ is a dummy variable for having less than $\overline{Exp}$ years of experience.

[25] The cutoff restriction is implemented by recoding experience as follows:

$$Exp_{jt} = \begin{cases} Exp_{jt} \text{ if } Exp_{jt} \leq \overline{Exp} \\ \overline{Exp} \text{ if } Exp_{jt} > \overline{Exp} \end{cases}$$

Results are similar with other cutoff levels, but this specification is preferred because teachers with more than ten years of experience teach about half of the students in the school-year cells in my data. Results are also similar with other polynomial specifications of $f\left(Exp_{jt}\right)$, but while the cubic term appears to be important in at least one subject area, quartic or higher order terms do not.

[26] Split-level classrooms refer to classes where students of adjacent grades are placed in the same classroom. This arrangement was used in district B, albeit infrequently, to help balance class sizes.

serially correlated within students over time, standard errors are clustered at the student level.[27]

As one might expect, students perform lower than their own average in years when they are subsequently held back—between .26 and .41 standard deviations on the nationally standardized scales. In the following year, when repeating a grade, students perform significantly lower than their average in the Math Computation subject area, but otherwise score similarly to their average.[28]

I find that classroom specific variables are not important predictors of test scores in these districts. Students in split-level classrooms, both above and below the split, do not perform significantly differently than they do in regular classrooms. Class size has a statistically insignificant effect on student test scores in all four subject areas, and the signs of the point estimates are split evenly between positive and negative values. In other regressions (not shown), I check for non-linear effects of class size by including its square and cube. I also try interacting class size with a dummy for being Black or Hispanic, since Krueger (1999) and Rivkin et al. (2001) find that minorities may be more sensitive to class size effects. I do not find statistically significant effects of class size in any of these specifications.

The average past performance of students' classmates also seems to have no discernible effect on test scores. In other specifications (not shown), I find linear and non-linear transformations of classmates' previous achievement or classmates' demographic characteristics

[27] Measurement error in test scores is heteroskedastic by construction. Since tests are geared toward measuring achievement at a particular grade and time, e.g. spring of 3rd grade, the test is less accurate for students who find the test very difficult or very easy.

[28] These estimates may reflect the influence of many factors associated with being held back or repeating a grade. Students who are held back may have had difficulties stemming from problems at home, an illness, etc. Likewise, when they repeat a grade, they may be getting more support from parents or may be working harder in order not to fail again, in addition to seeing material for a second time.

are also not significant predictors of students' own achievement.[29] Though it is quite difficult to know how peer effects operate *a priori*, there are many reasons to think that measures of past achievement and observable characteristics would be good proxies for peer effects.[30] Using past achievement also helps avoid the reflection problem in estimating contemporaneous peer influences (Manski 1993).

The insignificance of classroom characteristics in these regressions may be viewed as somewhat surprising, given the recent literature on these issues and evidence from some studies of teacher effects (Hanushek 1972, Summers and Wolfe 1979). However, these estimates should not be interpreted as causal, since I am not making an effort to credibly identify the effects of these variables from exogenous variation; I am including them as controls so that I can be certain that differences in teacher fixed effects are not driven by differences in these factors.

The use of past achievement as a control variable forces me to drop a substantial fraction of observations from my analysis—an entire grade and year—and does not help to explain test scores. I therefore remove this control variable and present results on teacher fixed effects and experience effects from regressions of test scores that include a larger number of students and teachers. Estimated effects for the time-varying student controls and other classroom factors are quite similar to those cited above, and are shown in table 2.[31]

The joint statistical significance of teacher fixed effects in these regressions is measured

[29] In particular, I also tried including the variance of classmates' test scores, and the number (or proportion) of a students' classmates who: 1) had previous scores one standard deviation above/below the mean, 2) were classified students 3) were enrolled in ESL, 4) were held back or repeating a grade, 5) were female, 6) were Black or Hispanic. I also tried various combinations and interactions of these factors.

[30] For instance, high achieving students may share knowledge with their classmates, low performing students may disrupt instruction, dispersion of achievement may make teaching more difficult, etc.

[31] The only notable difference is that, in Reading Comprehension, the negative effect of being placed in the lower half of a split-level classroom is now statistically significant.

BEST COPY AVAILABLE

by an F-test, and the F-statistics and their associated p-values are shown in Panel I of table 3. They indicate that teachers are highly significant predictors of achievement in all four subject areas, with p-values below .001. In order to be sure that outlying observations on transient teachers do not drive these results, I repeat these tests using only teachers observed in at least three years. P-values for this more selective test are lower in all subject areas.[32]

To express the magnitude of teacher fixed effects, I calculate the difference between the median teacher and those at various percentiles of the fixed effect distribution.[33] These calculations are shown in Panel II of table 3. Differences between the 75th and 25th percentile teachers in reading and math scores are about 5.5 and 6.5 NCE points, respectively, or about .26 and .31 standard deviations on the national achievement distribution.[34] If teacher effects are normally distributed, the estimates above imply that a one standard deviation change in teacher quality would change student test scores by about .20 and .24 standard deviations in reading and math, respectively. Transient teachers do not drive these results either; repeating these calculations using only teachers observed in at least three years gives similar magnitudes.

The difference between high and low quality teachers is given in terms of nationally standardized exam scores and is thus easily interpretable. However, it is difficult to know how the distribution of teacher quality in these districts compares to the distribution of quality among broader groups of teachers, for example, statewide or nationwide. Nevertheless, salaries, geographic amenities, and other factors that affect districts' abilities to attract teachers vary

[32] P-values for tests on teacher effects in the regressions that included classmates' previous test scores are all also below .001, both for all teachers and teachers observed at least three times.

[33] Though there are alternative ways of expressing variation in teacher quality, this method is simple and transparent. It is quite similar to that used by Bertrand and Schoar (2001) to characterize the magnitude of CEO fixed effects on firm outcomes.

[34] Comparing percentiles at different parts of the distribution gives similar results.

to a much greater degree at the state or national level. This suggests that variation in quality within groups of teachers at broader geographical levels may be considerably larger, and that my estimates of the importance of teachers may be conservative. The controls for school-year effects may also lead me to underestimate the magnitude of variation in teacher quality, since any variation in average teacher quality across school-year cells is taken up by these controls.[35]

To analyze experience effects, I plot point estimates and 95% confidence intervals for the function $f(Exp_{jt})$ in figures 2 and 3. These results provide substantial evidence that teaching experience improves reading test scores. Ten years of teaching experience is expected to raise both Vocabulary and Reading Comprehension test scores by about 4 NCE points or .2 standard deviations (figure 2). However, the path of these gains is quite different between the two subject areas. In line with the identifying assumption, the function for Vocabulary scores exhibits positive and declining marginal returns, and gains approach zero as experience approaches the cutoff point.

Marginal returns to experience exhibit much slower declines for Reading Comprehension, and suggest that my identification assumption may be violated in this case.[36] However, if returns to experience were positive after the cutoff, as it appears they might be, the experience function I estimate would be biased downward, because estimated school-year effects would be biased to rise over time. Thus, these results may provide a conservative estimate of the impact of teaching experience on Reading Comprehension test scores.

[35] F-tests of the joint significance of school-year effects show them to be important predictors of test scores in all four subject areas.

[36] The hypothesis that gains are zero near the cutoff cannot be rejected and the cubic term is negative, but the functional form of $f(Exp_{jt})$ appears fairly linear.

There is little evidence of gains from experience for the two math subjects (figure 3). While the first few years of teaching experience appear to raise scores significantly in Math Computation (about .1 standard deviations), subsequent years of experience appear to lower test scores, though standard errors are too large to conclude anything definitive about these trends. Math Concepts scores do not seem to be raised significantly by teaching experience at any point.

Estimates of experience effects should not be affected by any correlation between teachers' fixed effects and their propensity to remain teaching in these districts. However, if teachers who stay were selected based on their gains from experience, this identification strategy would lead to biased estimates of the expected experience effects for all teachers. While the direction of this potential bias is unclear, these estimates should be interpreted as the expected gains from experience for teachers who stay in these districts.[37]

I check the sensitivity of these results to the set of identifying assumptions by comparing them with estimates under two other sets of restrictions. In the first case, I assume that year effects are zero and include grade effects and school-grade interactions. This change in specification does not change the results except to increase the estimated impact of teaching experience. In the second case, I assume that student fixed characteristics are uncorrelated with teacher assignment, so that student fixed effects can be omitted, and all interactions between school, grade, and year can be included. This change produces larger estimated impacts of teacher fixed effects and teaching experience than those presented above.

[37] Teachers who improve greatly may be more likely to remain if they have gained more firm-specific or occupation-specific human capital, if the district administration is more likely to reappoint them, or if their probability of eventually being offered tenure has increased. On the other hand, if teachers tend to leave after a particularly bad year, and the cause of that poor performance is not persistent, then there may be a negative correlation between expected gains and the probability of staying.

4.1 Naïve Estimates of Experience Effects

Previous studies have relied on variation across teachers to identify experience effects, and are susceptible to bias from correlation between experience levels and other teacher characteristics that affect student achievement. This type of correlation could arise for many reasons: less effective teachers may be less likely to get reappointed, more effective teachers may be more likely to move to higher paying occupations, teacher quality may differ by cohort, etc.

To view these correlations, figure 4 plots averages of the estimated teacher fixed effects by years of experience, from zero to ten years. For ease of exposition, the average fixed effect for teachers with no experience is normalized to zero. There is a clear negative relation between teacher fixed effects and experience in the Vocabulary subject area, suggesting that estimates of experience effects that do not condition on teacher fixed effects would be much smaller for this test. Trends in the other subject areas are less stark—a small negative relation in Reading Comprehension and Math Concepts, and a small positive relation in Math Computation.

$$(4)\ A_{isjt} = \alpha_i + \gamma X_{it} + \mu M_j + f\left(Exp_{jt}\right) D_{Exp_{jt} \leq \overline{Exp}} + f\left(\overline{Exp}\right) D_{Exp_{jt} > \overline{Exp}} + \eta C_{jt} + \pi_{st} + \varepsilon_{isjt}$$

Results for experience effects that do not control for teacher fixed effects come from estimation of equation 4, where $f\left(Exp_{jt}\right)$ is a cubic and $\overline{Exp}$ is ten years.[38] In lieu of teacher fixed effects is a dummy for whether or not a teacher has a masters degree (M_j). Students' test scores are not significantly higher on average with teachers who have masters

[38] In order to make these estimates comparable to those above, experience effects are estimated only for teachers with experience at or below the cutoff. To implement this, I interact the cubic in experience with a dummy for having ten or less years of experience, and include a dummy variable for having more than ten years of experience.

degrees, and on Reading Comprehension tests they are significantly lower by about .02 standard deviations (see table 4).[39]

Figures 5 and 6 show the point estimates for experience effects from this specification.[40] As predicted, estimated returns to experience are much lower for Vocabulary test scores. They are also lower for Reading Comprehension and Math Concepts, and there is little evidence of statistically significant returns to experience in any test subject. As mentioned above, there are many reasons why experience and fixed teacher quality might be correlated, and the correlations shown in figure 4 cannot be generalized to other school districts. However, these findings provide clear evidence that using variation in student performance across teachers to measure gains from experience is likely to give misleading results.

4.2 Correlation of Teacher Quality Across Subjects

It is quite possible that a teacher is better at teaching one subject than another, and this variation in skill might be important for policy decisions. For example, if the quality of teachers' mathematics instruction was inversely related to the quality of reading instruction, then exchanging teachers between students would have an ambiguous effect on student outcomes, and having teachers specialize in teaching one subject might be more efficient. I briefly examine this question by looking at the pairwise correlations between teachers' fixed effects across subjects, shown in table 5. There are positive correlations between all tests, although correlations between Vocabulary and other subject areas are considerably smaller (.16 to .32) than among the other three subject areas (.46 to .67). There is little indication

[39] 40% of teachers in district A and 28% of teachers in district B have a masters degree.

[40] For ease of exposition 95% confidence intervals for these estimates, and the point estimates with controls for fixed effects are also shown.

that teachers who are better at mathematics instruction are worse at reading instruction or vice versa.[41]

Sampling error may bias measures of the correlation of teacher fixed effects across subjects, but the direction of the bias is unclear *a priori*. Errors that are common across subjects will lead to upward bias, and errors that are independent across subjects will lead to downward bias. If the true correlation between subjects is the same for all teachers in this sample, I can gain some insight into the direction of bias by recalculating the correlations using only teachers observed with at least three classrooms, since sampling error is smaller for this subsample. Pairwise correlations among this group of teachers are between .05 and .1 higher in all subject areas, indicating that sampling error is likely to have biased down the correlations shown in table 5.[42]

4.3 Variance Decomposition

To give an idea of the potential scope of teachers' impact on the overall distribution of scores, I estimate upper and lower bounds on the proportion of test score variance accounted for by teacher fixed effects and experience effects. This also serves to demonstrate the potential

[41] It is also possible that some teachers are better at teaching certain types of students than others. If this were true, then there might be efficiency gains through active matching of students and teachers. In contrast, if the 'good' teachers are equally good for everyone, then the matching of students and teachers probably has more to do with equity than efficiency.

To examine this issue, I estimate quantile regressions at the 25th, 50th, and 75th quantiles. These regressions are of the same form as that used to estimate equation 3, but do not include student fixed effects. (Including student fixed effects requires too much computational power. Even without student fixed effects, the estimated variance-covariance matrix of the estimators must be obtained via bootstrapping, and this can take weeks.) I find teacher fixed effects are significant predictors of test scores in all of these regressions. They are also positively correlated: correlation coefficients between the 25th and 75th quantiles for the same subject area range from .50 to .79.

[42] To truly correct for sampling error in these calculations, one would simultaneously estimate teacher effects on all four subject areas in a multiple equation regression framework, and locate the corresponding error variance estimates in the variance-covariance matrix. Since the direction of bias is likely to be downward, and these findings are only an extension to the main results above, I do not pursue this strategy.

scope of policies targeted at improving teacher quality. However, my data come from only two districts (and they are quite similar in many respects), so it would be naïve to draw conclusions from these results about how variation in teacher quality across districts might explain variation in achievement.

The upper bound estimate of the variance accounted for by teachers is the adjusted R^2 from a linear regression of test scores on teacher fixed effects and experience effects. The lower bound estimate is the increase in the adjusted R^2 when teacher fixed effects and experience effects are added to a regression specification that contains dummies for students who are retained or repeat grades, student fixed effects, and school-year effects.[43] For comparison, I also estimate lower and upper bounds in this same way for the school-year effects and the student level effects (i.e., fixed effects and the controls for being retained and repeating a grade). Table 6 shows these results. Across subject areas, the upper bound estimates range from 5.0-6.4% for teacher effects, 2.7-6.1% for school-year effects, and 59-68% for student fixed effects. The lower bound estimates range from 1.1-2.8% for teacher effects, .4% to 2.3% for school-year effects, and 57-64% for student effects.

The lower bound estimates of test score variance accounted for by teacher effects may seem small. However, when thinking about the role of policies, one should keep in mind that explaining the total variance in test scores with policy-relevant factors is probably impossible. Idiosyncratic factors and natural variation in cognitive ability among students are surely beyond policymakers' control. Moreover, policymakers often avoid intervention in the home, and household factors may play a large role in determining test score outcomes.

[43] I omit classroom characteristics from this part of my analysis because they do not have significant predictive power for test scores.

A better characterization may be to calculate the proportion of "policy-relevant" test score variance accounted for by teachers. An estimate of policy-relevant variance can be found by taking the fraction of test score variance due to measurement error–say .10–and the lower bound estimate of the fraction of test score variance attributed to student-level variables–.57 to .64–and subtracting their sum from 1.[44] Using the estimates in table 6, I find differences among teachers explain proportions of policy-relevant test score variance ranging from lower bounds of 4-9% to upper bounds of 16-23%.

5 Conclusion

The empirical evidence in this paper suggests that raising teacher quality may be a key instrument in improving student outcomes. However, in an environment where many observable teacher characteristics are not related to teacher quality, policies that focus on recruiting and retaining teachers with particular credentials may be less effective than policies that reward teachers based on performance.

As measures of effective teaching, test scores are widely available, objective, and (though they may not capture all facets of what students learn in school) they are widely recognized as important indicators of achievement by educators, policymakers, and the public. A number of states have begun rewarding teachers with non-trivial bonuses based on the average test performance of students in their schools, but few areas (Cincinnati, Denver) have pursued

[44] A tenth of variance due to measurement error is a standard and perhaps conservative estimate. Standardized test makers publish reliability coefficients, which estimate the correlation of test-retest scores for the same student, and these usually are about .9 or slightly below. One minus this reliability coefficient is equivalent to the percentage of variance due to idiosyncratic factors, or what we call measurement error for simplicity. On the other hand, it is probably the case that some of the variance in test scores stemming from cognitive ability and household factors can be affected by education-based policy initiatives. For example, special education programs may increase the average test score performance of students with learning disabilities. Measuring the degree to which this is possible is clearly an extremely difficult exercise, and certainly beyond the scope of this paper.

programs that link individual teacher salaries to their own students' achievement. Recent studies of pay-for-performance incentives for teachers in Israel (Lavy 2002a, Lavy 2002b) indicate that both group- and individual-based incentives have positive effects on students' test scores, and that individual-based incentives may be more cost-effective.

Teacher evaluations may also present a simple and potentially important indicator of teacher quality. There is already substantial evidence that principals' opinions of teacher effectiveness are highly correlated with student test scores (Murnane 1975, Armor et al. 1976), and while evaluations introduce an element of subjectivity, they may also reflect valuable aspects of teaching other than improving test performance.

However, efforts to improve the quality of public school teachers face some difficult hurdles, the most daunting of which is the growing shortage of teachers. Hussar (1998) estimated the demand for newly hired teachers between 1998 and 2008 at 2.4 million—a staggering figure, given that there were only about 2.8 million teachers in the U.S. during the 1999-2000 school year.[45] Underlying this prediction is the fact that the fraction of teachers nearing retirement age has been growing steadily over the past two decades and continues to do so. In 1978, 25.7% of elementary and secondary public school teachers were over the age of 45; by 1998 that figure was 47.8%.

There is also evidence that the supply of highly skilled teachers has declined. A recent study by Corcoran et al. (2002) shows that females with very high test scores who graduated

[45] Notably, this prediction does not take into account possible reductions in class size, which would considerably increase the need for new teachers. Even if lowering class size has a significant beneficial effect on student achievement, it will certainly cause a temporary drop in average experience levels, and may lower long run teacher quality if new teachers are of lower quality than current teachers. Moreover, the impact of class size reduction may vary by district, since wealthy districts may fill their increased demand for new teachers with the highest quality teachers from poorer areas. Jepsen and Rivkin (2002) provide evidence that this type of shifting in teacher quality took place after class size reduction legislation was enacted in California.

high school in the early 1980s were much less likely to enter teaching than those from earlier cohorts. One reason for this change may be that the opportunities outside of teaching for highly skilled females have improved. Indeed, the average income of female teachers, relative to college-educated women in other professions, has declined substantially over this time period.[46] Although recent evidence indicates women who were once full-time teachers usually do not leave the education profession for a job that pays more money (Scafadi et al. 2002), there may be many women (and men) who would make excellent teachers, but choose not to teach for monetary reasons.

Given this set of circumstances, it is clear that much research is still needed on how high quality teachers may be identified, recruited, and retained. Seeking out and compensating teachers solely on the basis of education and experience (above the first few years) is unlikely to yield large increases in teacher quality, though currently this is common practice. Finding alternative sources of information on teacher quality may be crucial to the creation of effective policies to raise student achievement.

References

[1] Abowd, John M and Francis Kramarz, "The Analysis of Labor Markets Using Matched Employer-Employee Data" Handbook of Labor Economics. Volume 3B. Ashenfelter, Orley Card, David, eds., Handbooks in Economics, vol. 5. Amsterdam; New York and Oxford: Elsevier Science, North-Holland. p 2629-2710. 1999.

[2] Angrist, Joshua and Kevin Lang, "How Important Are Classroom Peer Effects? Evidence from Boston's Metco Program," Working Paper, July, 2002.

[3] Angrist, Joshua and Victor Lavy, "Using Maimonides' Rule to Estimate the Effect of Class Size of Scholastic Achievement," Quarterly Journal of Economics, May, 1999.

[4] Armor, David, et al., "Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools," RAND Publication, August, 1976.

[5] Bertrand, M. and A. Schoar, "Managing with Style: The Effect of Managers on Firm Policies", Working Paper, University of Chicago and MIT, April, 2002.

[46] See Hanushek and Rivkin (1997).

[6] Corcoran, Sean, William Evans and Robert Schwab, "Changing Labor Market Opportunities for Women and the Quality of Teachers 1957-1992," Working Paper, August, 2002.

[7] Figlio, David N., "Teacher Salaries and Teacher Quality," Economics Letters, August, 1997.

[8] Hanushek, Eric A., "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data," American Economic Review, May, 1971.

[9] Hanushek, Eric A., "The Economics of Schooling: Production and Efficiency in Public Schools," Journal of Economic Literature, September, 1986.

[10] _____, "The Trade-Off between Child Quantity and Quality," Journal of Political Economy, February 1992.

[11] Hanushek, Eric A. and Steven G. Rivkin, "Understanding the Twentieth-Century Growth in U.S. School Spending," Journal of Human Resources, Winter, 1997.

[12] _____, "New Evidence About *Brown v. Board of Education*: The Complex Effects of School Racial Composition on Achievement," Working Paper, October, 2002.

[13] Hoxby, Caroline, "Peer Effects in the Classroom: Learning from Gender and Race Variation," NBER Working Paper 7867, August, 2000.

[14] _____, "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," Quarterly Journal of Economics, November, 2000.

[15] Hussar, William J., "Predicting the Need for Newly Hired Teachers in the United States to 2008-09," Research and Development Report, National Center for Educational Statistics, August, 1999.

[16] Jepsen, Christopher and Steven Rivkin, "What is the Tradeoff Between Smaller Classes and Teacher Quality?," NBER Working Paper 9205, September, 2002.

[17] Kane, Thomas and Douglas Staiger, "Improving School Accountability Measures," NBER Working Paper 8156, March, 2001.

[18] Krueger, Alan, "Experimental Estimates of Education Production Functions," Quarterly Journal of Economics, May, 1999.

[19] Lavy, Victor, "Paying for Performance: The Effect of Teachers' Financial Incentives on Students' Scholastic Outcomes," Working Paper, July, 2002.

[20] Lavy, Victor, "Evaluating the Effect of Teacher Performance Incentives on Students' Achievements," Journal of Political Economy, December, 2002.

[21] Manski, Charles, "Identification of Endogenous Social Effects: The Reflection Problem," Review of Economic Studies, July, 1993.

[22] Murnane, Richard, The Impact of School Resources on the Learning of Inner City Children. Cambridge, MA: Ballinger, 1975.

[23] Rivkin, Steven G., Eric A. Hanushek and John F. Kain, "Teachers, Schools, and Academic Achievement," Working Paper, April, 2001.

[24] Scafidi, Benjamin, David Sjoquist and Todd R. Stinebrickner, "Where Do Teachers Go?," Working Paper, October, 2002.

[25] Summers, Anita and Barbara Wolfe, "Do Schools Make a Difference?," American Economic Review, September, 1977.

A Tests for Systematic Classroom Assignment

To test for systematic differences in the groups of students assigned to particular teachers (i.e., tracking), I test if current classrooms are significant predictors of past test scores. To do so, I calculate the residuals from a regression of past test scores on school-year-grade dummies, regress these residuals on classroom dummies, and test the significance of variation in past test scores across classrooms using a joint F-test on these dummy variables. I only look at variation within school-year-grade cells because administrators can only change the classroom to which they assign students, not the school, year or grade. Table A.1 shows, by district, the F-statistics and p-values for these tests in each of the four subject areas. All of the p-values are close to one, substantiating administrators' claims that there was no systematic classroom assignment based on ability/achievement.

I also examine how students are mixed from year to year as they progress to higher grades, i.e., if administrators tend to keep the same groups of students together for successive years. This type of systematic classroom assignment would not be captured by differences in past achievement across classrooms. I examine this issue through calculation of dissimilarity indices, commonly used to measure spatial segregation (e.g. of racial groups in neighborhoods within a city). One can see the intuition for using this measure by asking: are students in a particular school-grade-year cell 'segregated' across current classrooms by their previous classroom? If one considers a school-grade-year cell like a city, a classroom like a neighborhood, and a student's previous classroom like a racial group, the issues are clearly parallel.

To indicate what dissimilarity indices would look like with random assignment, I generate data where students from four 'classrooms' of 20 students each are randomly placed into four new 'classrooms' of 20 students each—this is fairly representative of the school-year-grade cells in my data. Dissimilarity indices from this monte carlo exercise are located predominantly between .1 and .3. Figure A.1 shows, by district, the actual proportion of school-grade-year dissimilarity indices falling between zero and .1, .1 and .2, etc. A large majority of cells have indices between .1 and .3, giving strong evidence that the mixing of classmates from year to year in these districts is similar to random assignment.[47]

[47] Though indices decrease with the number of students in each classroom and increase with the number of classrooms, but large changes in the parameters I use (e.g., 100 students per classroom or 20 classrooms per school-grade cell) are needed to radically change the results. Also, a tiny fraction of school-grade-year cells in district B have indices above .6. This is driven by the small number of classrooms in district B that are 'split-level', i.e. they have students from adjacent grades placed in the same classroom. It is obvious when looking at the data that many of the students placed in the lower grade of a split-level classroom remain with that teacher the following year if that teacher is assigned a split-level classroom.

Table 1: Estimated Effects of Student Characteristics and Classroom Characteristics on Test Scores

	Reading Vocabulary	Reading Comprehension	Math Computation	Math Concepts
Held Back	-5.546	-5.569	-8.533	-8.060
	(1.830)**	(1.879)**	(2.736)**	(2.120)**
Repeating Grade	0.841	2.032	-4.296	-1.118
	(1.853)	(2.041)	(2.147)*	(1.926)
Class Size	0.045	-0.129	-0.081	0.099
	(0.080)	(0.078)	(0.107)	(0.077)
Classmates' Average Previous Test Score	-0.004	0.023	0.051	-0.009
	(0.032)	(0.030)	(0.039)	(0.030)
Split-level Classroom	0.414	-0.715	-0.883	0.194
	(0.735)	(0.632)	(0.805)	(0.734)
Below Split in Split-level Classroom	0.139	-1.318	-0.098	-1.503
	(0.765)	(0.722)	(0.843)	(0.810)
Observations	17409	20506	18266	23289
R-squared	0.82	0.82	0.81	0.85

Test scores are expressed on a Normal Curve Equivalent scale; one standard deviation on this scale is 21 points. All regressions include teacher and student fixed effects, a cubic in experience, and school-year effects. Standard errors (in parentheses) are clustered by pupil. * significant at 5%; ** significant at 1%

Table 2: Estimated Effects of Student Characteristics and Classroom Characteristics on Test Scores

	Reading Vocabulary	Reading Comprehension	Math Computation	Math Concepts
Held Back	-6.323	-6.088	-9.508	-8.621
	(1.858)**	(1.854)**	(2.437)**	(1.993)**
Repeating Grade	0.969	2.062	-0.086	-1.117
	(1.864)	(1.989)	(2.154)	(1.826)
Class Size	0.046	-0.100	0.102	0.108
	(0.068)	(0.064)	(0.077)	(0.062)
Split-level Classroom	0.848	0.225	-0.496	0.124
	(0.605)	(0.536)	(0.613)	(0.597)
Below Split in Split-level Classroom	-0.128	-1.364	0.084	-1.065
	(0.687)	(0.622)*	(0.704)	(0.699)
Observations	22335	26012	25006	29312
R-squared	0.82	0.82	0.79	0.83

Test scores are expressed on a Normal Curve Equivalent scale; one standard deviation on this scale is 21 points. All regressions include teacher and student fixed effects, a cubic in experience, and school-year effects. Standard errors (in parentheses) are clustered by pupil. * significant at 5%; ** significant at 1%

Table 3: Significance and Magnitude of Teacher Fixed Effects

	Reading Vocabulary	Reading Comprehension	Math Computation	Math Concepts
Panel I: Significance of Teacher Fixed Effects †		F-statistic	(P-Value)	
Reading Vocabulary		2.82	(<0.001)	
Reading Comprehension		2.07	(<0.001)	
Math Computation		3.65	(<0.001)	
Math Concepts		5.54	(<0.001)	
Magnitude of Teacher Fixed Effects ††	*Difference between median and __th percentile*			
	10th	*25th*	*75th*	*90th*
Reading Vocabulary	-5.59	-2.78	3.15	5.53
Reading Comprehension	-4.48	-2.36	3.23	5.81
Math Computation	-6.57	-3.56	3.19	7.93
Math Concepts	-7.51	-3.35	3.19	6.01

Regressions include controls for being held back or repeating a grade, class size, being in a split-level classroom and being in the lower half of a split-level classroom, student fixed effects, school-year effects, and experience effects.

† F-test is on the joint significance of teacher dummy variables to predict test scores in the linear regression.

†† Differences across teacher effects are given in terms of points on a Normal Curve Equivalent scale; one standard deviation on this scale is 21 points.

Table 4: Omitting Teacher Fixed Effects from Test Score Regressions

	Reading Vocabulary	Reading Comprehension	Math Computation	Math Concepts
Held Back	-7.970	-6.978	-10.453	-9.708
	(2.072)**	(2.133)**	(2.815)**	(2.154)**
Repeating Grade	-0.088	2.019	-1.355	-0.522
	(1.936)	(2.364)	(2.541)	(2.137)
Class Size	-0.019	-0.051	0.025	0.085
	(0.061)	(0.056)	(0.070)	(0.055)
Split-level Classroom	0.712	0.425	-0.349	1.087
	(0.501)	(0.465)	(0.521)	(0.538)*
Below Split in Split-level Classroom	-0.542	-1.772	0.114	-1.340
	(0.620)	(0.571)**	(0.659)	(0.661)*
Teacher Has Masters Degree	-0.165	-0.475	-0.054	0.189
	(0.247)	(0.226)*	(0.284)	(0.233)
Observations	21780	25354	24460	28657
R-squared	0.82	0.81	0.77	0.81

Test scores are expressed on a Normal Curve Equivalent scale; one standard deviation on this scale is 21 points. All regressions include teacher and student fixed effects, a cubic in experience, and school-year effects. Standard errors (in parentheses) are clustered by pupil. * significant at 5%; ** significant at 1%

Table 5: Correlation of Teacher Fixed Effect Estimates Across Subject Area Tests

	Reading Vocabulary	Reading Comprehension	Math Computation	Math Concepts
Reading Vocabulary	1.00			
Reading Comprehension	0.27	1.00		
Math Computation	0.16	0.46	1.00	
Math Concepts	0.32	0.58	0.67	1.00

Note: These are the pairwise correlations of teacher fixed effects across subjects. The teacher fixed effects used to calculate these correlations are estimated in regressions of test scores that include controls for students who are retained or repeat a grade, class size, being in a split-level classroom and being in the lower half of a split-level classroom, student fixed effects, a cubic in experience, and school-year effects.

Table 6: Test Score Variance Decomposition

	Upper Bound R-sq[1]	Lower Bound R-sq[2]	*Base R-sq*[2]
Teacher Fixed Effects and Experience			
Reading Vocabulary	0.050	0.018	*0.690*
Reading Comprehension	0.051	0.011	*0.691*
Mathematics Computation	0.052	0.028	*0.619*
Mathematics Concepts	0.064	0.025	*0.700*
School-Year Effects			
Reading Vocabulary	0.034	0.009	*0.699*
Reading Comprehension	0.039	0.004	*0.698*
Mathematics Computation	0.027	0.015	*0.632*
Mathematics Concepts	0.061	0.023	*0.703*
Student-Level Effects			
Reading Vocabulary	0.676	0.643	*0.065*
Reading Comprehension	0.683	0.641	*0.061*
Mathematics Computation	0.595	0.575	*0.073*
Mathematics Concepts	0.658	0.624	*0.102*

Notes: 1. Upper bound estimates are the adjusted R^2 from a regression of test scores on just the factor in question: school year effects, teacher dummy variables and a cubic in experience, or student fixed effects and controls for students who are retained or repeat a grade.
2. Lower bound estimates are the increase in adjusted R^2 from adding one of the sets of factors to a regression of test scores that included the other two sets of factors as controls. The adjusted R^2 from this latter regression is the Base R^2, shown in the third column.

Figure 1: Distribution of School Districts' Test Scores Relative to National Distribution

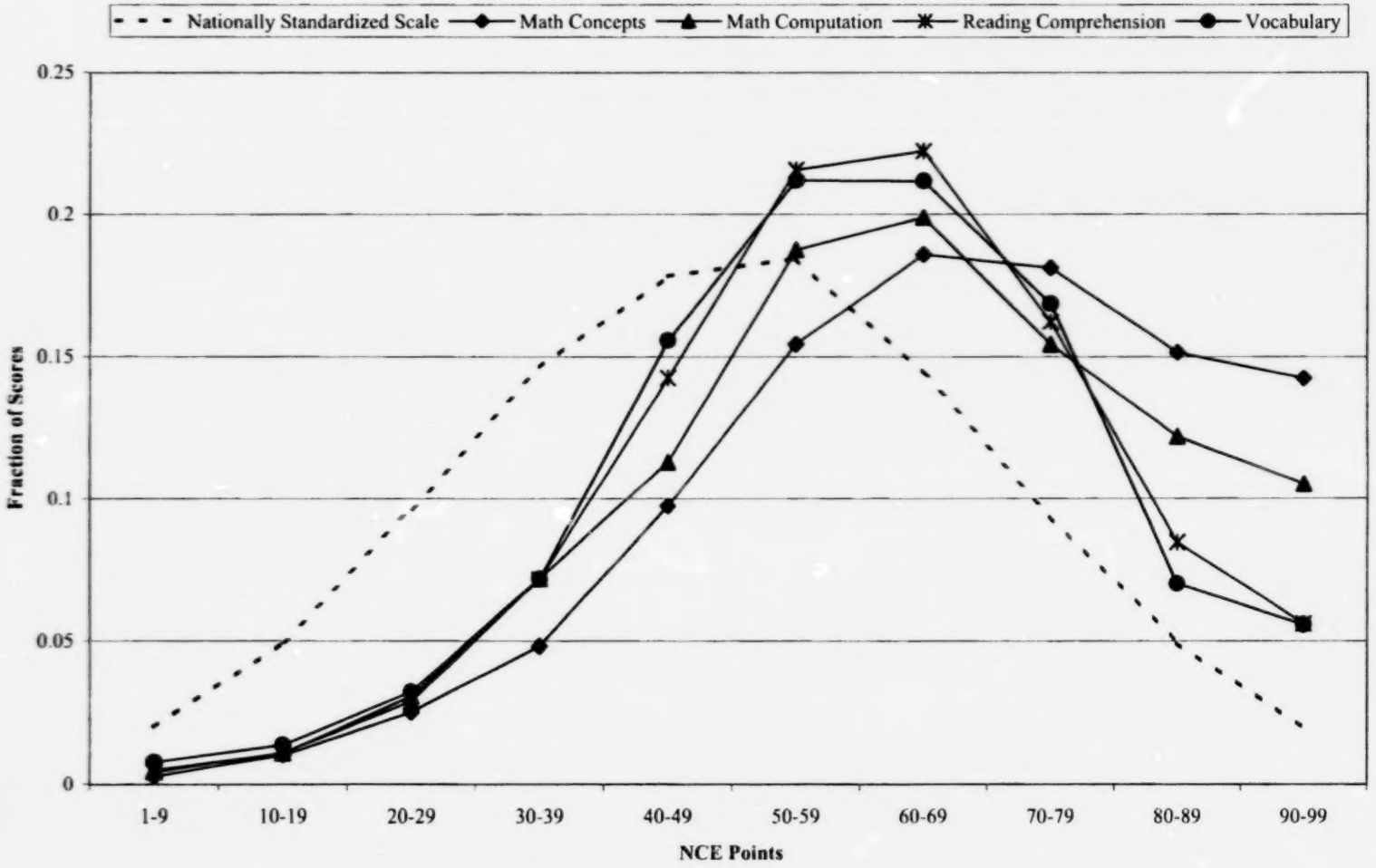
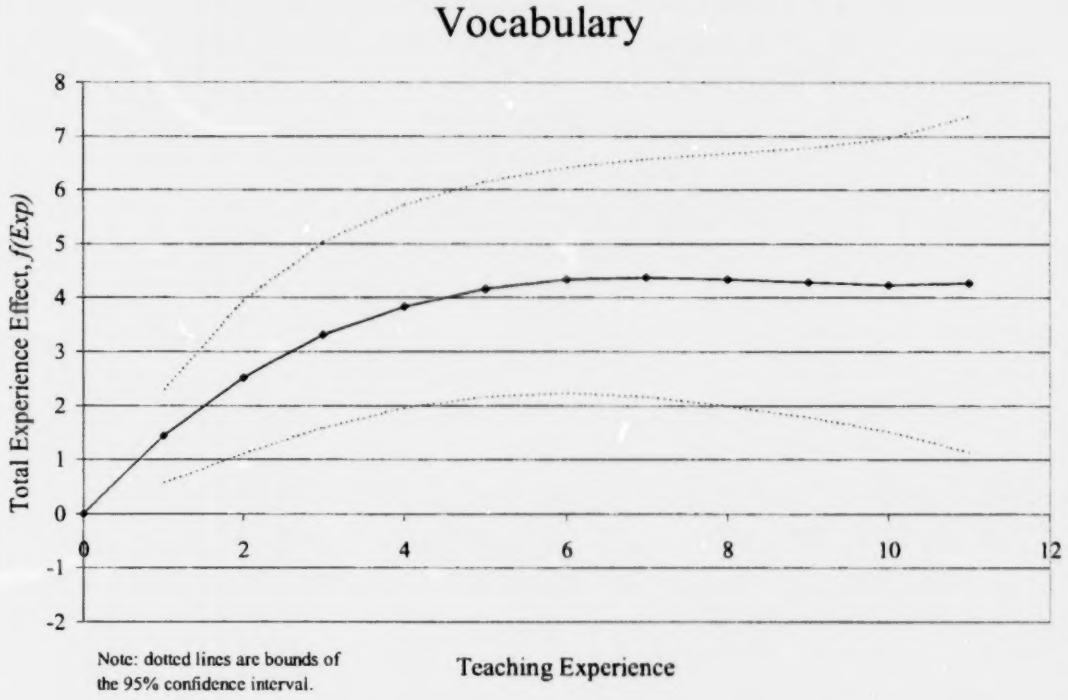


Figure 2: The Effect of Teacher Experience on Reading Achievement, Controlling for Fixed Teacher Quality



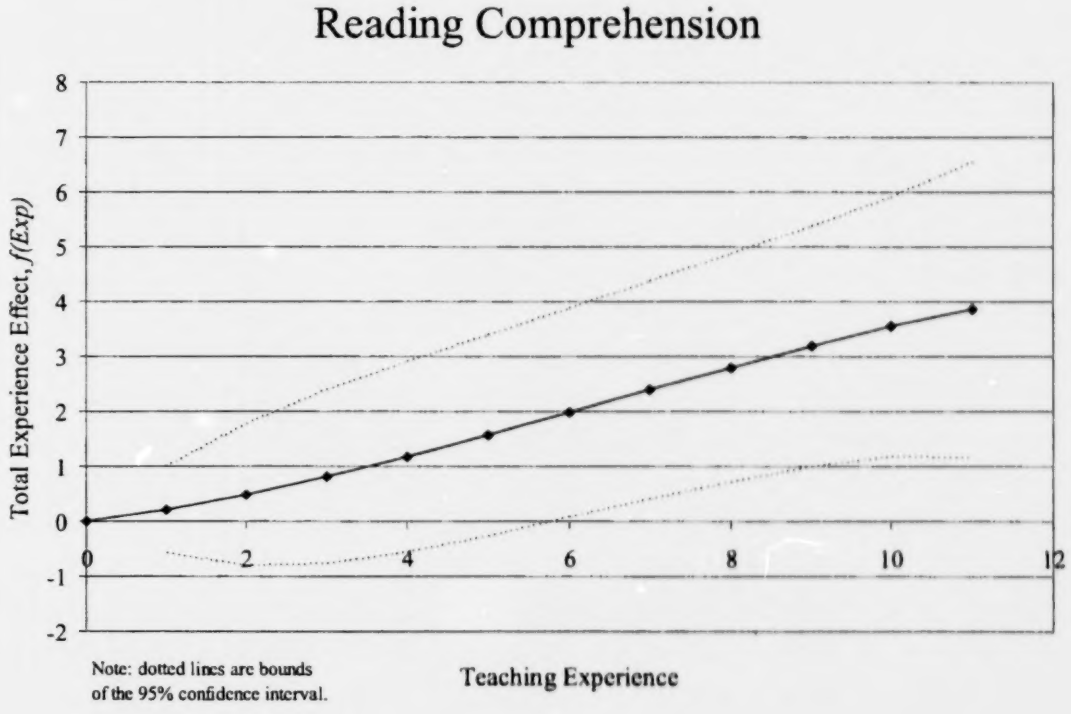
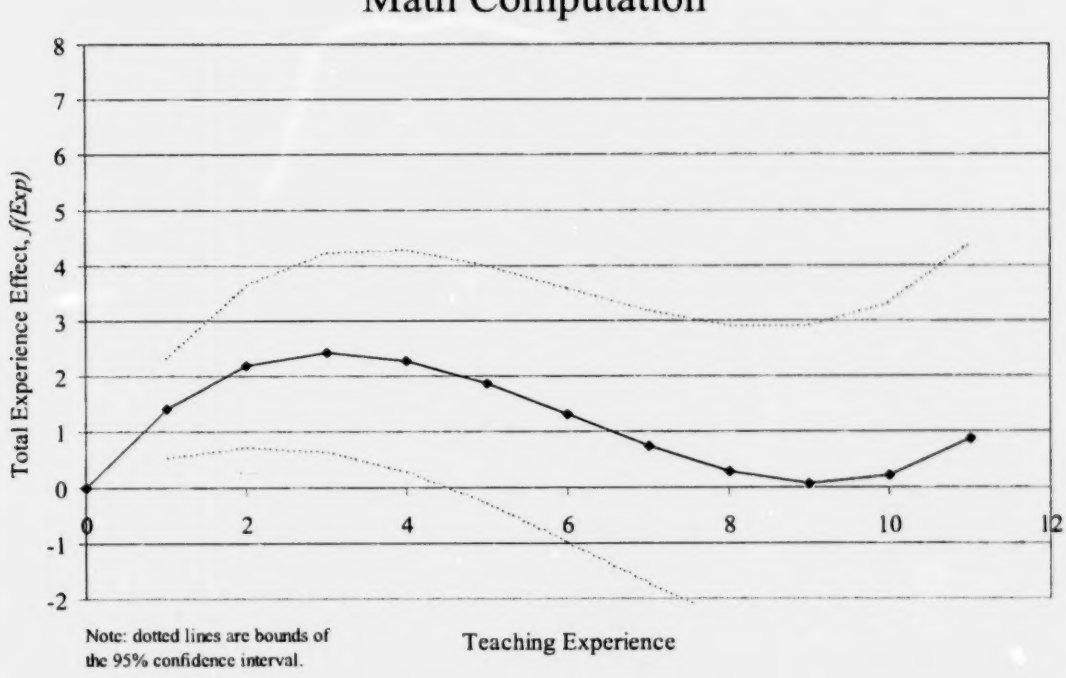


Figure 3: The Effect of Teacher Experience on Math Achievement, Controlling for Fixed Teacher Quality



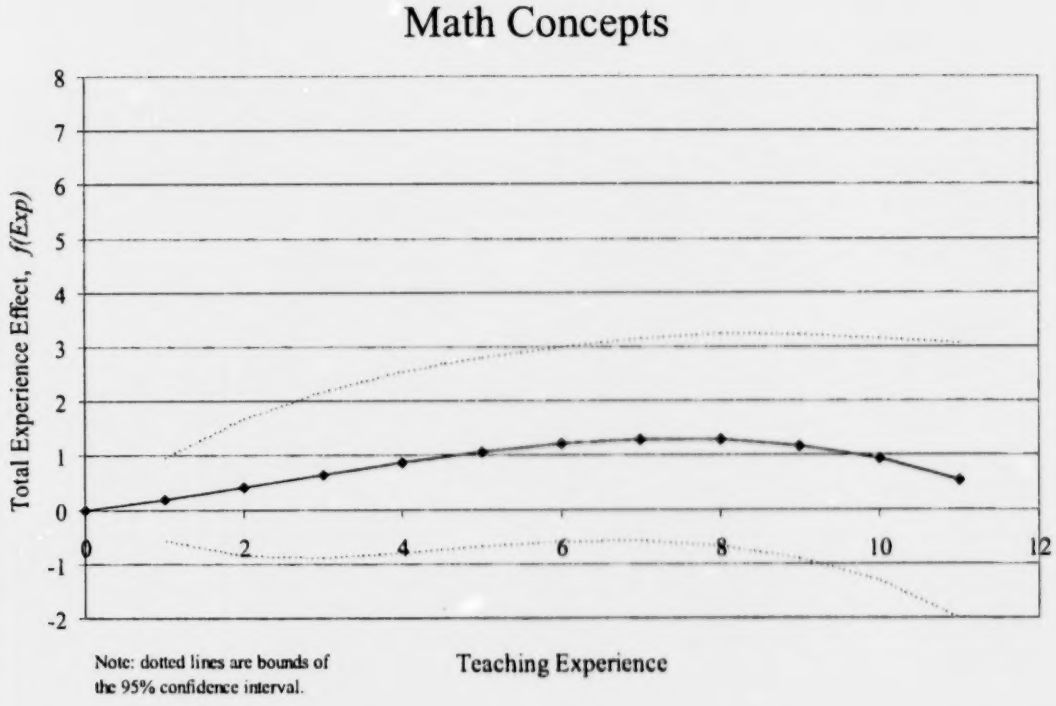
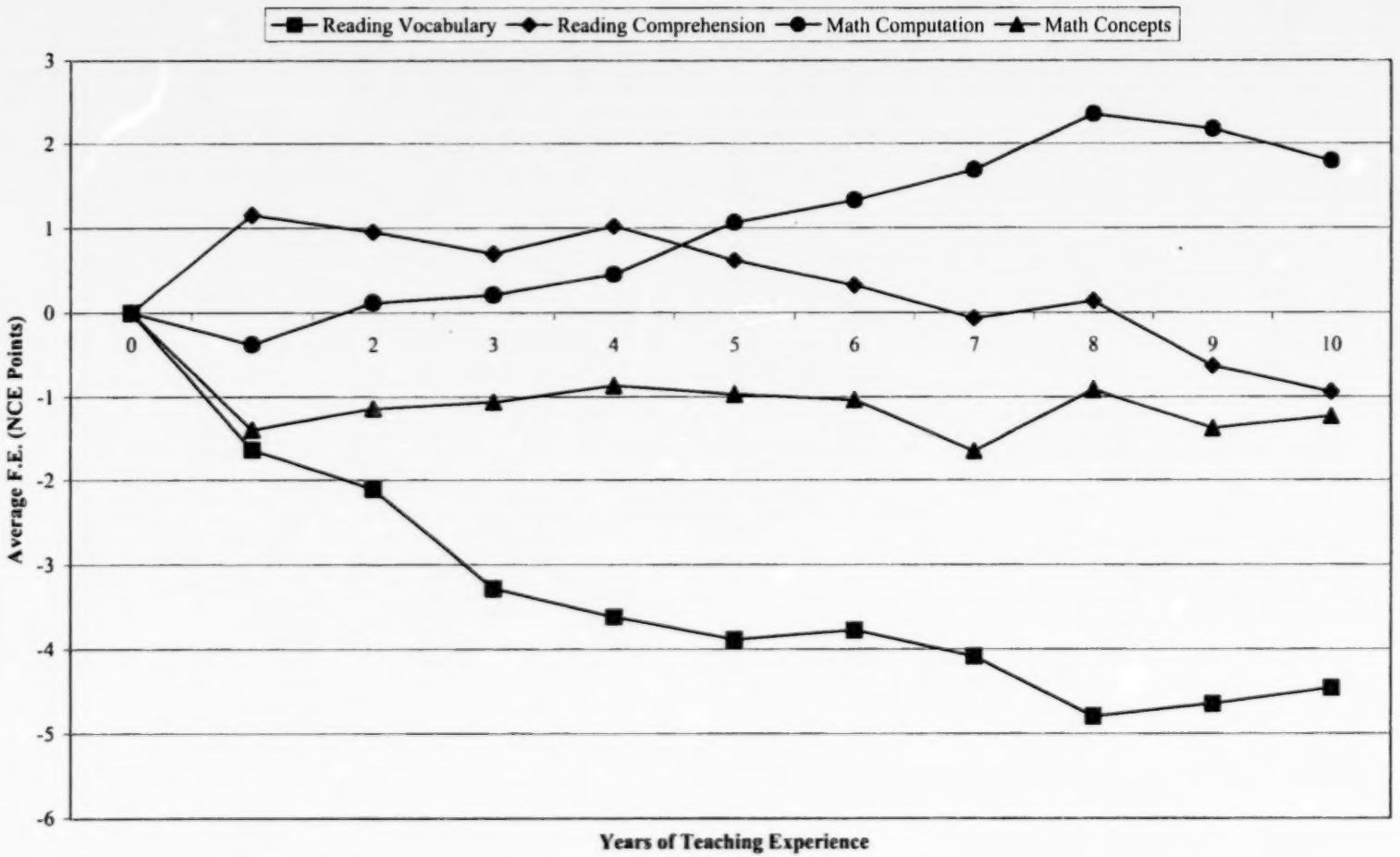
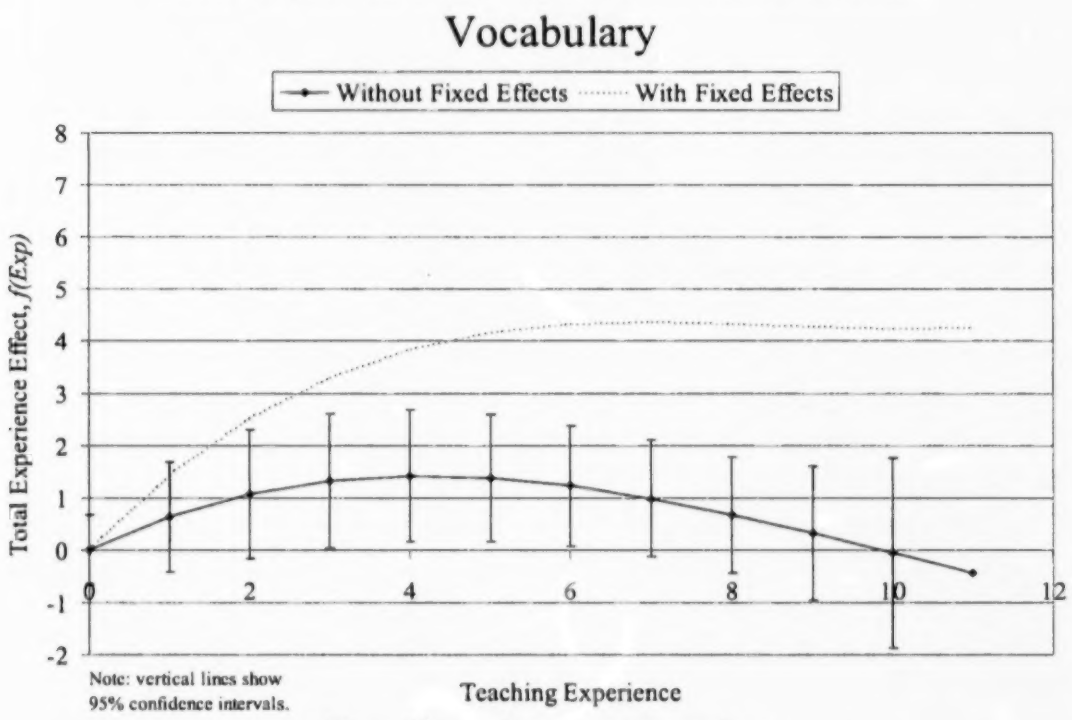


Figure 4: Average Teacher Fixed Effect by Experience Level



Note: Estimates of teacher fixed effects are taken from regressions that include controls for students who are retained or repeat grades, class size, being in a split-level classroom and being in the lower half of a split-level classroom, student fixed effects, a cubic in experience, and school-year effects.

Figure 5: Teacher Experience Effects on Reading Achievement, The Impact of Omitting Teacher Fixed Effects



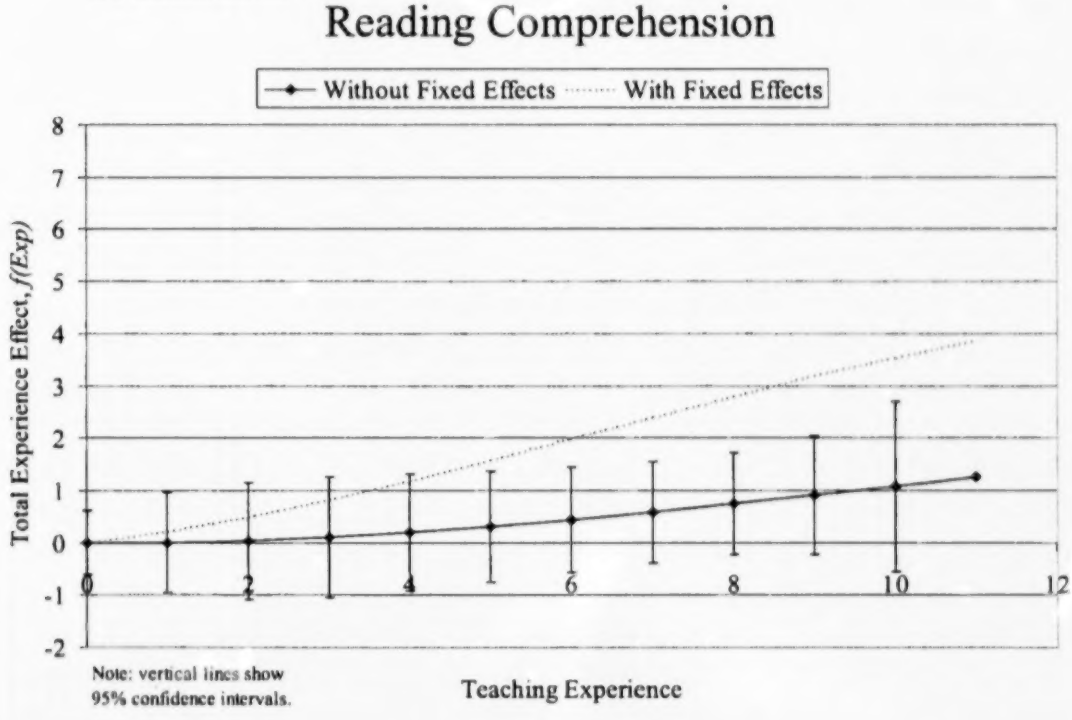
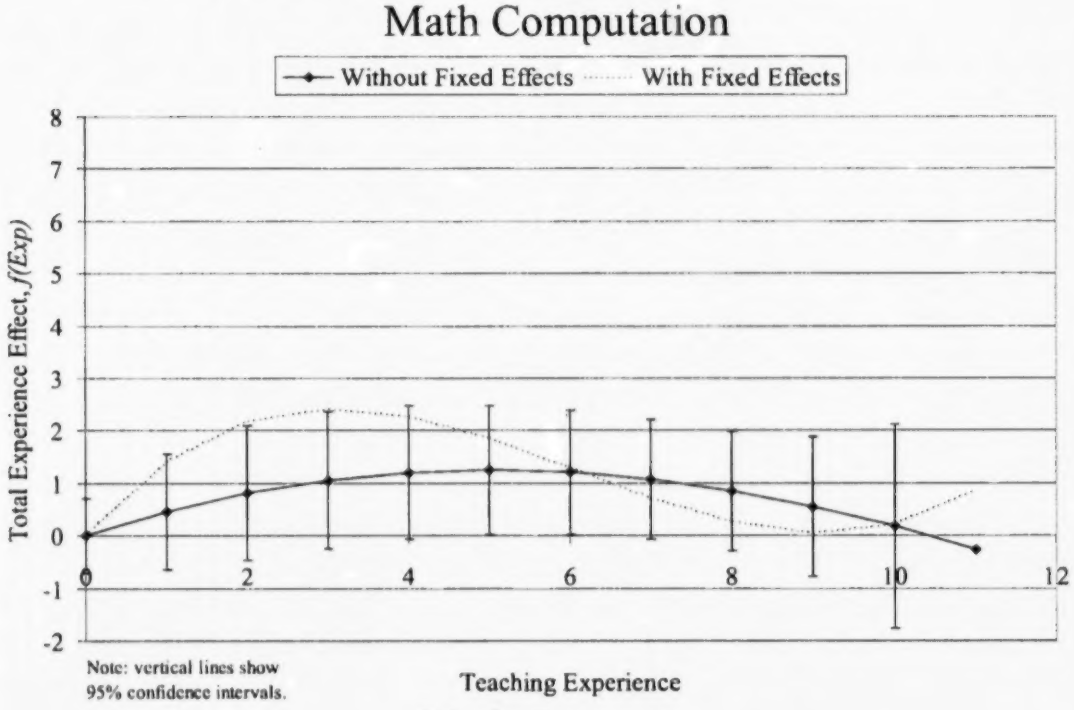


Figure 6: Teacher Experience Effects on Math Achievement, The Impact of Omitting Teacher Fixed Effects



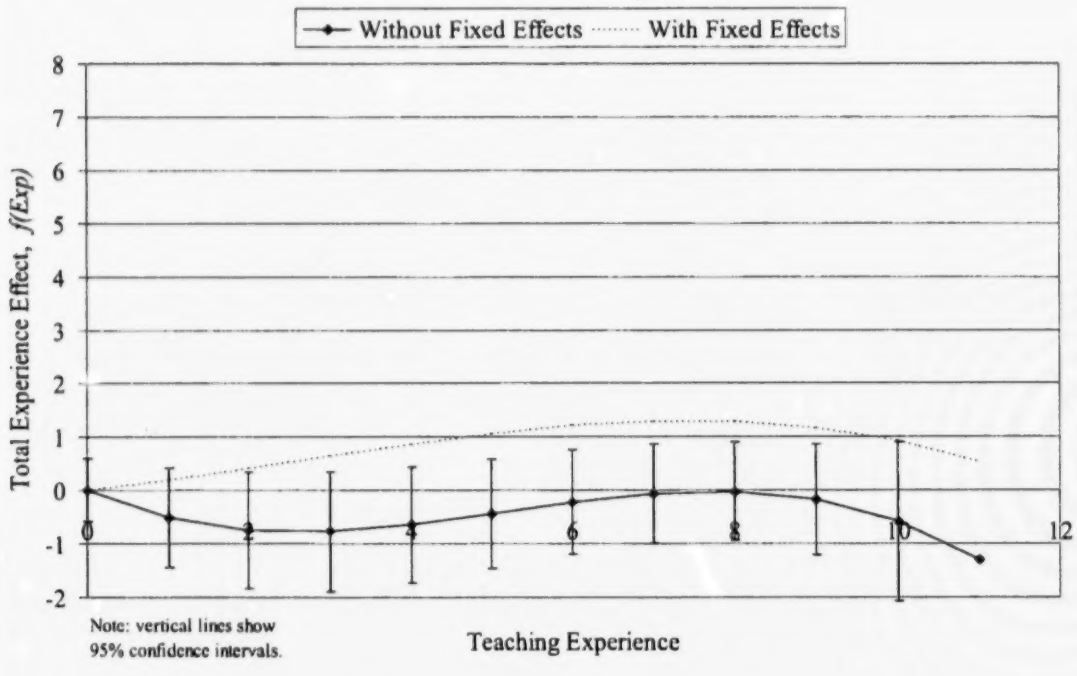
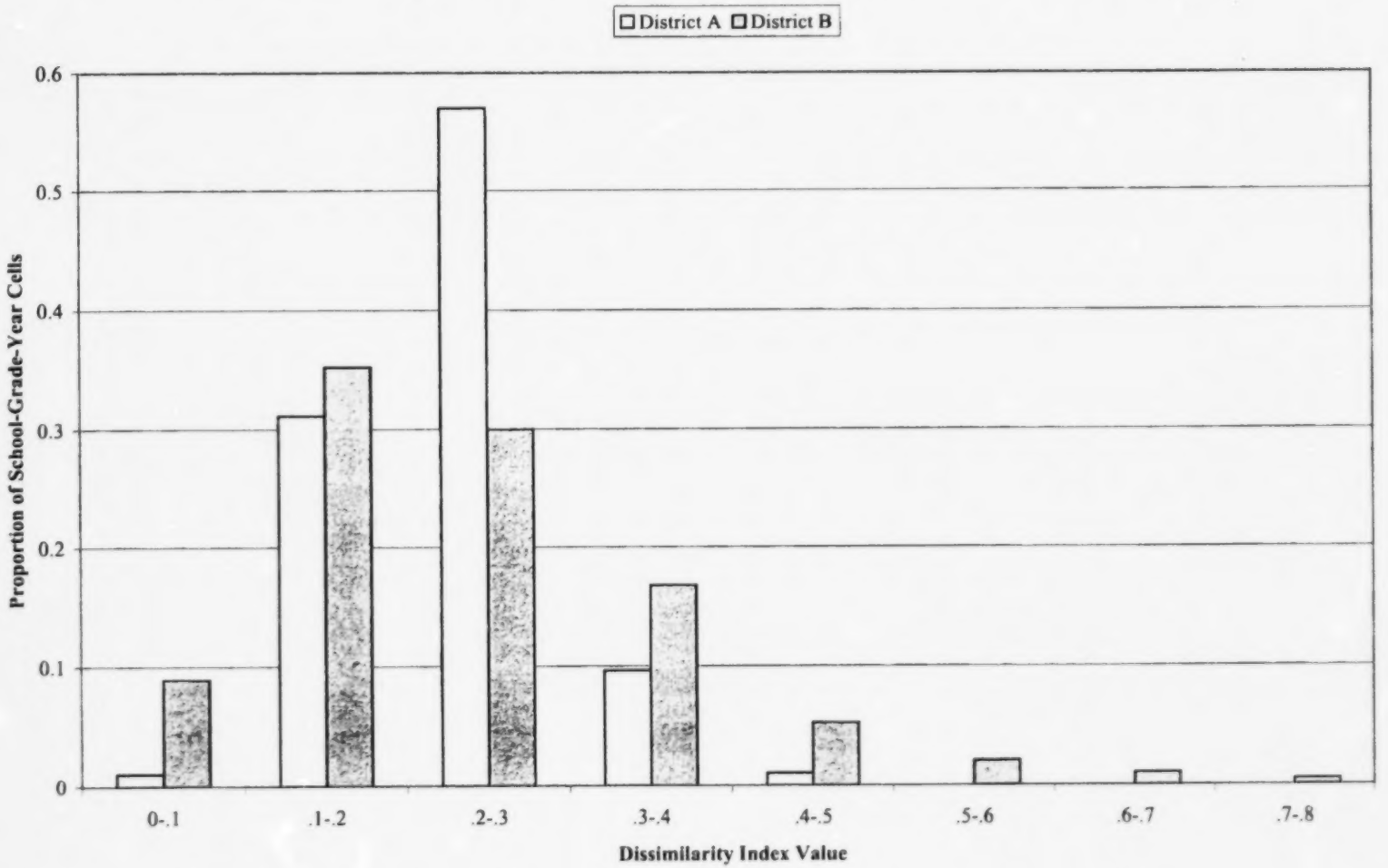


Table A.1: Statistical Tests for Tracking by District and Test

	District A		District B	
	F-statistic	P-value	F-statistic	P-value
Reading Vocabulary	0.74	1.00	0.88	0.97
Reading Comprehension	0.77	1.00	0.91	0.90
Math Computation	0.77	1.00	0.90	0.95
Math Concepts	0.74	1.00	0.94	0.85

Notes: F-tests are on the joint significance of classroom dummies to predict past test scores within school-year-grade cells.

Figure A.1: Dissimilarity Indices by School-Grade-Year Cell (Segregation by Previous Classroom)



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

UD035615

I. DOCUMENT IDENTIFICATION:

Title: "The Impact of Individual Teachers on Student Achievement: Evidence From Panel Data"	
Author(s): Jonah E. Rockoff	
Corporate Source: Not Applicable, Document is a Working Paper	Publication Date: March, 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to each document.

If permission is granted to reproduce and disseminate the identified documents, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

SAMPLE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

SAMPLE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

SAMPLE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate these documents as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Jonah Rockoff	Printed Name/Position/Title: Jonah Rockoff, Ph.D. Candidate, Harvard University
Organization/Address: Harvard, 1080 Massachusetts Ave., Cambridge, MA 02138-5398	Telephone: 617-813-1217 / FAX: 617-868-3742
	E-Mail Address: rockoff@fas.harvard.edu / Date: April 22, 2003